(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－53394

(43)公開日　平成11年(1999)２月26日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/403	３５０Ｃ
17/21		15/20	５７０Ｎ
		15/403	３４０Ａ

審査請求　未請求　請求項の数11　ＦＤ　（全 11 頁）

(21)出願番号　特願平9－219298

(22)出願日　平成９年(1997)７月29日

(71)出願人　390024350
株式会社ジャストシステム
徳島県徳島市沖浜東３－46

(72)発明者　野村　直之
徳島県徳島市沖浜東３丁目46番地　株式会社ジャストシステム内

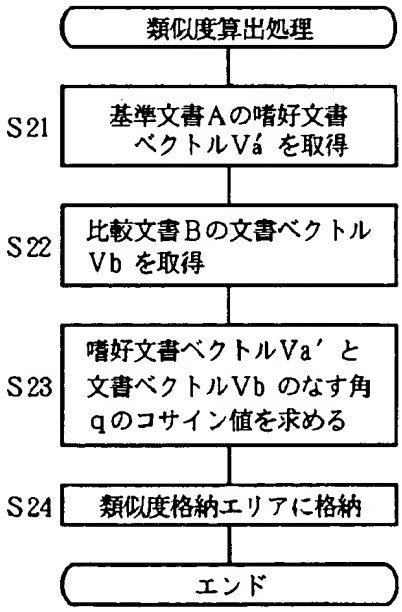
(74)代理人　弁理士　川井　隆　（外１名）

(54)【発明の名称】　文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法

(57)【要約】

【課題】　ユーザーの嗜好を踏まえた文書処理を行うことのできる文書処理装置、文書処理プログラムを記憶した記憶媒体、及び文書処理方法を提供すること。

【解決手段】　過去の処理文書中の出現頻度等から処理重要語句（キーワード）を取得し、処理重要語句の処理文書中の出現頻度等からユーザー全体の嗜好を表すＧＰベクトルを取得する。また、類似度を取得する基準となる基準文書Ａにおける処理重要語句の重要度を取得し、この重要度を要素とする文書ベクトルＶａを取得し、ＧＰベクトルと文書ベクトルＶａの各要素を掛け合わせて嗜好文書ベクトルＶ’ａを得る。この嗜好文書ベクトルには、ユーザーの嗜好が反映されている。嗜好文書ベクトルＶ’ａと他の文書Ｂの文書ベクトルＶｂ，とでなす角ｑのコサイン値ｃｏｓ（ｑ）を、基準文書Ａと他の文書Ｂとの類似度とする。この類似度が大きいほど、他の文書はユーザーの嗜好に近いものとなる。



1

【特許請求の範囲】

【請求項１】　ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルを取得する嗜好ベクトル取得手段と、
文書を取得する文書取得手段と、
前記文書取得手段により取得された文書を特徴付ける文書ベクトルを取得する文書ベクトル取得手段と、
前記文書ベクトル取得手段により取得された前記文書ベクトルを前記嗜好ベクトルによりシフトさせるシフト手段とを具備することを特徴とする文書処理装置。

【請求項２】　前記嗜好ベクトル取得手段は、前記ユーザーが作成した文書又はアクセスした文書に対する複数の文書ベクトルから前記嗜好ベクトルを作成することを特徴とする請求項１に記載の文書処理装置。

【請求項３】　複数のユーザーと、複数の前記ユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列とし、前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列を取得するＧＰ行列取得手段を備え、
前記嗜好ベクトル取得手段は、前記ＧＰ行列により前記嗜好ベクトルを取得することを特徴とする請求項１または請求項２に記載の文書処理装置。

【請求項４】　前記ＧＰ行列取得手段は、
ユーザーが過去に処理した文書と該文書のキーワードとのうちの一方を行、他方を列とし、前記文書におけるキーワードの重要度を要素値とする文書－キーワード行列と、
ユーザーが過去に処理した文書とユーザーとのうちの一方を行、他方を列とし、各ユーザーの前記文書の処理回数を要素とする文書－ユーザー行列と、からＧＰ行列を取得することを特徴とする請求項３に記載の文書処理装置。

【請求項５】　所定の文書の文書ベクトルを嗜好ベクトルによりシフトさせた嗜好文書ベクトルと、他の文書の文書ベクトルから、前記所定の文書と前記他の文書との類似度を算出する類似度算出手段を具備することを特徴とする請求項１から請求項４のうちのいずれか１の請求項に記載の文書処理装置。

【請求項６】　ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルを取得する嗜好ベクトル取得機能と、
文書を取得する文書取得機能と、
前記文書取得機能により取得された文書を特徴付ける文書ベクトルを取得する文書ベクトル取得機能と、
前記文書ベクトル取得機能により取得された前記文書ベクトルを前記嗜好ベクトルによりシフトさせるシフト機能とをコンピュータに実現させるためのコンピュータ読み取り可能な文書処理プログラムが記憶された記憶媒体。

【請求項７】　前記嗜好ベクトル取得機能は、前記ユーザーが作成した文書又はアクセスした文書に対する複数の文書ベクトルから前記嗜好ベクトルを作成することを特徴とする請求項６に記載の文書処理プログラムが記憶された記憶媒体。

2

【請求項８】　複数のユーザーと、複数の前記ユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列とし、前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列を取得するＧＰ行列取得機能を備え、
前記嗜好ベクトル取得機能は、前記ＧＰ行列により前記嗜好ベクトルを取得することを特徴とする請求項６または請求項７に記載の文書処理プログラムが記憶された記憶媒体。

【請求項９】　前記ＧＰ行列取得機能は、
ユーザーが過去に処理した文書と該文書のキーワードとのうちの一方を行、他方を列とし、前記文書におけるキーワードの重要度を要素値とする文書－キーワード行列と、
ユーザーが過去に処理した文書とユーザーとのうちの一方を行、他方を列とし、各ユーザーの前記文書の処理回数を要素とする文書－ユーザー行列と、からＧＰ行列を取得することを特徴とする請求項８に記載の文書処理プログラムが記憶された記憶媒体。

【請求項１０】　所定の文書の文書ベクトルを嗜好ベクトルによりシフトさせた嗜好文書ベクトルと、他の文書の文書ベクトルから、前記所定の文書と前記他の文書との類似度を算出する類似度算出機能を具備することを特徴とする請求項６から請求項９のうちのいずれか１の請求項に記載の文書処理プログラムが記憶された記憶媒体。

【請求項１１】　ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルと、文書とを取得し、
前記文書を特徴付ける文書ベクトルを取得し、
前記文書ベクトルを前記嗜好ベクトルによりシフトさせることを特徴とする文書処理方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法に関し、更に詳細には、利用目的等のユーザーの嗜好を踏まえた類似文書の検索に関する。

【０００２】

【従来の技術】従来の文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法による文書処理においては、複数の文書を検索・分類するために、それぞれの文書について当該文書を特徴付ける文書ベクトルを取得し、この文書ベクトルから文書間の類似度を算出することが行われている。

【０００３】

【発明が解決しようとする課題】しかし、同一の文書でも、例えば営業用や技術資料用等の利用目的その他のユーザーの嗜好が異なると、重要部位等に差異が生じる。そして、上述のような文書ベクトルを用いた文書処理によって文書の検索や分類をしても、ユーザーの嗜好を踏まえた処理は行うことができないため、この様な嗜好を加味した上での文書処理を行うことのできる文書処理装置や文書処理プログラムが記憶された記憶媒体、文書処理方法が望まれていた。

【０００４】本発明は、上述のような課題を解決するためになされたもので、ユーザーの嗜好を踏まえた文書処理を行うことのできる文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法を提供することを目的とする。

【０００５】

【課題を解決するための手段】請求項１に記載の発明は、図９に示すように、ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルを取得する嗜好ベクトル取得手段１０１と、文書を取得する文書取得手段１０２と、前記文書取得手段１０２により取得された文書を特徴付ける文書ベクトルを取得する文書ベクトル取得手段１０３と、前記文書ベクトル取得手段により取得された前記文書ベクトルを前記嗜好ベクトルによりシフトさせるシフト手段１０４とを具備する文書処理装置を提供することにより前記目的を達成するものである。請求項２に記載の発明は、図９に示すように、請求項１に記載の文書処理装置において、前記嗜好ベクトル取得手段１０１は、前記ユーザーが作成した文書又はアクセスした文書に対する複数の文書ベクトルから前記嗜好ベクトルを作成する文書処理装置を提供することにより前記目的を達成するものである。請求項３に記載の発明は、図１０に示すように、請求項１または請求項２に記載の文書処理装置において、複数のユーザーと、複数の前記ユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列とし、前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列を取得するＧＰ行列取得手段１０５を備え、前記嗜好ベクトル取得手段１０１は、前記ＧＰ行列により前記嗜好ベクトルを取得する文書処理装置を提供することにより前記目的を達成するものである。請求項４に記載の発明は、図１０に示すように、請求項３に記載の文書処理装置において、前記ＧＰ行列取得手段１０５は、ユーザーが過去に処理した文書と該文書のキーワードとのうちの一方を行、他方を列とし、前記文書におけるキーワードの重要度を要素値とする文書－キーワード行列と、ユーザーが過去に処理した文書とユーザーとのうちの一方を行、他方を列とし、各ユーザーの前記文書の処理回数を要素とする文書－ユーザー行列と、からＧＰ行列を取得する文書処理装置を提供することにより前記目的を達成するものである。請求項５に記載の発明は、図１１に示すように、請求項１から請求項４のうちのいずれか１の請求項に記載の文書処理装置において、所定の文書の文書ベクトルを嗜好ベクトルによりシフトさせた嗜好文書ベクトルと、他の文書の文書ベクトルから、前記所定の文書と前記他の文書との類似度を算出する類似度算出手段１０６を具備する文書処理装置を提供することにより前記目的を達成するものである。請求項６に記載の発明は、図１２に示すように、ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルを取得する嗜好ベクトル取得機能２０１と、文書を取得する文書取得機能２０２と、前記文書取得機能２０２により取得された文書を特徴付ける文書ベクトルを取得する文書ベクトル取得機能２０３と、前記文書ベクトル取得機能２０３により取得された前記文書ベクトルを前記嗜好ベクトルによりシフトさせるシフト機能２０４とをコンピュータに実現させるためのコンピュータ読み取り可能な文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成するものである。請求項７に記載の発明は、図１２に示すように、請求項６に記載の記憶媒体において、前記嗜好ベクトル取得機能２０１は、前記ユーザーが作成した文書又はアクセスした文書に対する複数の文書ベクトルから前記嗜好ベクトルを作成する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成するものである。請求項８に記載の発明は、図１３に示すように、請求項６または請求項７に記載の記憶媒体において、複数のユーザーと、複数の前記ユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列とし、前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列を取得するＧＰ行列取得機能２０５を備え、前記嗜好ベクトル取得機能２０１は、前記ＧＰ行列により前記嗜好ベクトルを取得する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成する。請求項９に記載の発明は、図１３に示すように、前記ＧＰ行列取得機能２０５は、ユーザーが過去に処理した文書と該文書のキーワードとのうちの一方を行、他方を列とし、前記文書におけるキーワードの重要度を要素値とする文書－キーワード行列と、ユーザーが過去に処理した文書とユーザーとのうちの一方を行、他方を列とし、各ユーザーの前記文書の処理回数を要素とする文書－ユーザー行列と、からＧＰ行列を取得する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成するものである。請求項１０に記載の発明は、図１４に示すように、請求項６から請求項９のうちのいずれか１の請求項に記載の記憶媒体において、所定の文書の文書ベクトルを嗜好ベクトルによりシフトさせた嗜好文書ベクトルと、他の文書の文書ベクトルから、前記所定の文書と前記他の文書との類似度を算出する類似度算出機能２０６を具備する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を

達成するものである。請求項１１に記載の発明は、図１５に示すように、ユーザーの嗜好を表す複数のキーワードに対する重要度を要素値とする嗜好ベクトルと文書とを取得３０１し、前記文書を特徴付ける文書ベクトルを取得３０２し、前記文書ベクトルを前記嗜好ベクトルによりシフト３０３させる文書処理方法を提供することにより前記目的を達成するものである。

【０００６】

【発明の実施の形態】以下、本発明の文書処理装置、文書処理プログラムが記憶された記憶媒体及び文書処理方法の好適な実施の形態について、図１から図８を参照して詳細に説明する。

（１）実施形態の概要

本実施形態では、過去の処理文書中の出現頻度等から処理重要語句（キーワード）ａ，ｂ，…を取得し、処理重要語句の処理文書中の出現頻度、各処理文書の処理回数、処理したメンバーの重要度から、ユーザー全体の嗜好を表すＧＰベクトルを取得する。また、類似度を取得する基準となる基準文書Ａにおける処理重要語句ａ，ｂ，…の重要度ｇ（ａ），ｇ（ｂ），…を取得し、重要度ｇ（ａ），ｇ（ｂ），…を要素とする文書ベクトルＶａを取得し、ＧＰベクトルと文書ベクトルＶａ，Ｖｂ，Ｖｃ，…の各要素を掛け合わせて嗜好文書ベクトルＶ’ａを得る。この嗜好文書ベクトルには、ユーザーの嗜好が反映されている。嗜好文書ベクトルＶ’ａと他の文書の文書ベクトルＶｂ，とでなす角ｑのコサイン値ｃｏｓ（ｑ）を、基準文書Ａと他の文書Ｂとの類似度とする。この類似度が大きいほど、他の文書はユーザーの嗜好に近いものとなる。

【０００７】（２）実施形態の詳細

図１は、本発明の文書処理装置の一実施形態であり、本発明の文書処理プログラムが記憶された記憶媒体の一実施形態の該プログラムが読み取られたコンピュータの構成を表したブロック図である。この図１に示すように、文書処理装置（コンピュータ）は、装置全体を制御するための制御部１１を備えている。この制御部１１には、データバス等のバスライン２１を介して、入力装置としてのキーボード１２やマウス１３、表示装置１４、印刷装置１５、記憶装置１６、記憶媒体駆動装置１７、通信制御装置１８、入出力Ｉ／Ｆ１９、及び文字認識装置２０が接続されている。制御部１１は、ＣＰＵ１１１、ＲＯＭ１１２、ＲＡＭ１１３を備えている。ＲＯＭ１１２は、ＣＰＵ１１１が各種制御や演算を行うための各種プログラムやデータが予め格納されたリードオンリーメモリである。

【０００８】ＲＡＭ１１３は、ＣＰＵ１１１にワーキングメモリとして使用されるランダムアクセスメモリである。このＲＡＭ１１３には、本実施形態による文書ベクトル取得処理を行うためのエリアとして、文書ベクトル取得の対象となる文書を格納する対象文書格納エリア１１３１、キーワード格納エリア１１３２、文書ベクトル格納エリア１１３４が確保され、また、嗜好文書ベクトル取得処理を行うためのエリアとして、行列格納エリア１１３５、嗜好文書ベクトル格納エリア１１３６、類似度格納エリア１１３７その他の各種エリアが確保されるようになっている。

【０００９】キーボード１２は、かな文字を入力するためのかなキーやテンキー、各種機能を実行するための機能キー、カーソルキー、等の各種キーが配置されている。マウス１３は、ポインティングデバイスであり、表示装置１４に表示されたキーやアイコン等を左クリックすることで対応する機能の指定を行う入力装置である。表示装置１４は、例えばＣＲＴや液晶ディスプレイ等が使用される。この表示装置１４には、文書ベクトルを取得する対象文書の内容や、本実施形態により算出された文書間の類似度、算出された類似度をもとに行った検索結果や分類結果等が表示されるようになっている。印刷装置１５は、表示装置１４に表示された文章や、記憶装置１６の文書データベース１６４に格納された文書等の印刷を行うためのものである。この印刷装置としては、レーザプリンタ、ドットプリンタ、インクジェットプリンタ、ページプリンタ、感熱式プリンタ、熱転写式プリンタ、等の各種印刷装置が使用される。

【００１０】記憶装置１６は、読み書き可能な記憶媒体と、その記憶媒体に対してプログラムやデータ等の各種情報を読み書きするための駆動装置で構成されている。この記憶装置１６に使用される記憶媒体としては、主としてハードディスクが使用されるが、後述の記憶媒体駆動装置１７で使用される各種記憶媒体のうちの読み書き可能な記憶媒体を使用するようにしてもよい。記憶装置１６は、仮名漢字変換辞書１６１、プログラム格納部１６２、文書データベース１６４、重要語データベース１６５、行列データベース１６８、文書ベクトルデータベース１６６、嗜好文書ベクトルデータベース１６７、図示しないその他の格納部（例えば、この記憶装置１６内に格納されているプログラムやデータ等をバックアップするための格納部）等を有している。プログラム格納部１６２には、本実施形態における嗜好文書ベクトル取得処理プログラム、類似度算出処理プログラム等の各種プログラムの他、仮名漢字変換辞書１６１を使用して入力された仮名文字列を漢字混り文に変換する仮名漢字変換プログラム等の各種プログラムが格納されている。

【００１１】文書データベース１６４には、仮名漢字変換プログラムにより作成された文書や、他の装置で作成されて記憶媒体駆動装置１７や通信制御装置１８から読み込まれた文書が格納される。この文書データベース１６４に格納される各文書の形式は特に限定されるものではなく、テキスト形式の文書、ＨＴＭＬ（Hyper TextMarkup Language）形式の文書、ＪＩＳ形式の文書等の各種形式の文書の格納が可能である。更にこの文書データ

ベース１６４には、文書を処理したユーザー（処理者）及びその処理回数が各文書に対応付けて格納されている。前記処理回数は、所定期間毎に値を０にリセットされる。重要語データベース１６５には、前記所定期間内に処理した処理文書から抽出された重要語句（処理重要語句）及びその重要度（処理重要度）が格納される。

【００１２】行列データベース１６８には、過去の所定期間に行われた文書処理の処理内容により取得される行列Ｇａ，Ｇｂ，Ｇｃが格納されている。文書ベクトルは、これらの行列Ｇａ，Ｇｂ，Ｇｃにより取得されるＧＰ（Group Personalize）行列をもとに、嗜好文書ベクトルに変換される。図２（ａ）～（ｃ）は、行列Ｇａ，Ｇｂ，Ｇｃを示す説明図である。

【００１３】行列Ｇａ（文書－キーワード行列）は、図２（ａ）に示すように、前記所処理重要語句を行に、同処理文書を列にとった行列であり、各要素は処理重要語句の処理重要度ｆ（ｘ）を表している。行列Ｇｂ（文書－ユーザー行列）は、図２（ｂ）に示すように、前記処理文書を行にとり、ユーザーのメンバーを列にとった行列であり、各要素は、メンバーが各文書を前記所定期間内に処理した回数となっている。行列Ｇｃは、図２（ｃ）に示すように、行および列がともにユーザーのメンバーそれぞれの重要度係数を示している。行列Ｇａ及び行列Ｇｂは所定期間ごとに書き換えられ、行列Ｇｃは操作者からの入力により適宜書き換えられる。

【００１４】文書ベクトルデータベース１６６、及び嗜好文書ベクトルデータベース１６７には、本実施形態において類似度を算出する基準となる基準文書、及び、該基準文書に対する類似度を比較する対象文書それぞれの文書ベクトル及び嗜好文書ベクトルが格納される。

【００１５】記憶媒体駆動装置１７は、ＣＰＵ１１１が外部の記憶媒体からコンピュータプログラムや文書を含むデータ等を読み込むための駆動装置である。記憶媒体に記憶されているコンピュータプログラムには、本実施形態の文書処理装置により実行される各種処理のためのプログラム、および、そこで使用される辞書、データ等も含まれる。ここで、記憶媒体とは、コンピュータプログラムやデータ等が記憶される記憶媒体をいい、具体的には、フロッピーディスク、ハードディスク、磁気テープ等の磁気記憶媒体、メモリチップやＩＣカード等の半導体記憶媒体、ＣＤ－ＲＯＭやＭＯ、ＰＤ（相変化書換型光ディスク）等の光学的に情報が読み取られる記憶媒体、紙カードや紙テープ等の用紙（および、用紙に相当する機能を持った媒体）を用いた記憶媒体、その他各種方法でコンピュータプログラム等が記憶される記憶媒体が含まれる。本実施形態の文書処理装置において使用される記憶媒体としては、主として、ＣＤ－ＲＯＭやフロッピーディスクが使用される。記憶媒体駆動装置１７は、これらの各種記憶媒体からコンピュータプログラムを読み込む他に、フロッピーディスクのような書き込み可能な記憶媒体に対してＲＡＭ１１３や記憶装置１６に格納されているデータ等を書き込むことが可能である。

【００１６】本実施形態の文書処理装置では、制御部１１のＣＰＵ１１１が、記憶媒体駆動装置１７にセットされた外部の記憶媒体からコンピュータプログラムを読み込んで、記憶装置１６の各部に格納（インストール）する。そして、本実施形態による類似度算出等の各種処理を実行する場合、記憶装置１６から該当プログラムをＲＡＭ１１３に読み込み、実行するようになっている。但し、記憶装置１６からではなく、記憶媒体駆動装置１７により外部の記憶媒体から直接ＲＡＭ１１３に読み込んで実行することも可能である。また、文書処理装置によっては、本実施形態の自動要約処理プログラム等を予めＲＯＭ１１２に記憶しておき、これをＣＰＵ１１１が実行するようにしてもよい。

【００１７】通信制御装置１８は、他のパーソナルコンピュータやワードプロセッサ等との間でテキスト形式やＨＴＭＬ形式等の各種形式の文書やビットマップデータ等の各種データの送受信を行うことができるようになっている。入出力Ｉ／Ｆ１９は、音声や音楽等の出力を行うスピーカ等の各種機器を接続するためのインターフェースである。文字認識装置２０は、用紙等に記載された文字をテキスト形式やＨＴＭＬ等の各種形式で認識する装置であり、イメージスキャナや文字認識プログラム等で構成されている。

【００１８】本実施形態では、キーボード１２の入力操作により作成した文書（ＲＡＭ１１３の所定格納エリアに格納）の他、外部で作成して所定の記憶媒体に格納した文書で記憶媒体駆動装置１７から読み込んだ文書、予め文書データベース１６４に格納されている文書、通信制御装置１８からダウンロードした文書、及び文字認識装置２０で文字認識した文書、等の各種文書を対象文書として取得することが可能である。

【００１９】次に、上述のような構成の文書処理装置による嗜好文書ベクトル取得処理及び類似度算出処理について図面を参照して説明する。

【００２０】本実施形態においては、所定期間毎に、該所定期間内に行われた文書処理の処理内容基づいて新たな処理重要語句及び処理重要度が取得され、行列データベース１６８内の行列Ｇａ及び行列Ｇｂが書き換えられる。

【００２１】図３は、行列Ｇａ，Ｇｂ書き換え処理の動作を表したフローチャートである。ＣＰＵ１１１は、所定期間内に処理された文書（処理文書）を文書データベース１６４から順次取得してＲＡＭ１１３の所定作業領域に格納し（ステップ１１）、各処理文書についての重要語句（処理重要語句）及びその重要度（処理重要度）を取得する（ステップ１２）。

【００２２】図４は処理重要語句・処理重要度取得処理の動作を表したフローチャートである。図４に示すよう

に、ＣＰＵ１１１は、文書データベース１６４から取得した処理文書について、各処理文書毎に形態素解析を行うことで自立語を抽出する（ステップ１２１）と共に、名詞句、複合名詞句等を含めた候補語（句）を処理文書から抽出する（ステップ１２２）。次に、抽出した候補語（句）の処理文書での出現頻度、評価関数から、各候補語（句）の処理重要度ｆ（ｘ）を取得する（ステップ１２３）。ここで、評価関数としては、例えば、所定の重要語句が予め指定されている場合にはその重要語句に対する重み付け、単語、名詞句、複合名詞句等の候補語（句）の種類による重み付け等が使用される。

【００２３】さらにＣＰＵ１１１は、取得した処理重要度ｆ（ｘ）の値をもとに候補語（句）から処理重要語句ａ，ｂ，…を取得し（ステップ１２４）、この処理重要語句ａ，ｂ，…及びその処理重要度ｆ（ａ），ｆ（ｂ），…を重要語データベース１６５に格納する（ステップ１２５）。すべての処理文書について、処理重要語句及びその処理重要度を取得すると、図４に示す行列Ｇａ，Ｇｂ書き換え処理ルーチンへリターンする。

【００２４】続いて、ＣＰＵ１１１は、行列データベース１６８の行列Ｇａを、前記処理重要語句ａ，ｂ，…を行に、前記所定期間の処理文書を列に、また処理重要度ｆ（ａ），ｆ（ｂ），…を各要素にとったものに書き換える（ステップ１３）。このとき、行列Ｇａの行数は、各処理文書の処理重要語句の和集合の数とし、各処理文書において含まれていない処理重要語句については、その処理重要度ｆ（ｘ）は０と定義される。

【００２５】例えば図２おいて、処理文書Ｂの処理重要語は「重要、重要語、重要度、…」、処理文書Ｃの処理重要語句は「重要、…、政治、…」であり、これらの処理重要語句に対応する処理重要度は、処理文書Ｂについては（　１，１８，１９，…）、処理文書Ｃについては（１８，…，２１，…）である。これに対して行列Ｇａにおいては、その行は「重要、重要語、重要度、…、政治、…」とし、両文書の列における要素値はつぎの通り定義される。

処理文書Ｂの列＝（　１，１８，１９，…，　０，…）、

処理文書Ｃの列＝（１８，　０，　０，…，２１，…）

【００２６】また、ＣＰＵ１１１は、文書データベース１６４から、各処理文書の処理回数を取得し（ステップ１４）、行列Ｇｂを、所定期間内の処理文書を行に、文書データベース１６４から取得した処理回数を各要素としたものに書き換えて（ステップ１５）、行列Ｇａ，Ｇｂ書き換え処理を終了する。

【００２７】図５は、本実施形態による文書嗜好ベクトル取得処理の動作を示すフローチャートである。嗜好文書ベクトル取得に際しては、嗜好文書ベクトル取得の対象となる文書（対象文書）を取得し、ＲＡＭ１１３の対象文書格納エリア１１３１に格納する（ステップ２１）。対象文書は、ユーザの指示に従ってＲＡＭ１１３、記憶装置１６の文書データベース１６４、記憶媒体駆動装置１７、または通信制御装置１８（パソコン通信、インターネット等の通信による場合）から取得する。

【００２８】次にＣＰＵ１１１は、対象文書中から行列Ｇａの処理重要語句を抽出する（ステップ２２）。次に、抽出した処理重要語句の対象文書中での出現頻度、評価関数等から、重要度ｇ（ｙ）を取得する（ステップ２３）。ここで、評価関数としては、例えば、処理重要語句に対する予め指定されている重み付け、単語、名詞句、複合名詞句等の候補語（句）の種類による重み付け等が使用される。

【００２９】そして、この処理重要語句ａ，ｂ，…の対象文書中における重要度ｇ（ａ），ｇ（ｂ），…を要素とする文書ベクトルＶを取得する（ステップ２４）。

【００３０】文書ベクトルＶを取得すると、ＣＰＵ１１１は、行列データベース１６８から行列Ｇａ，Ｇｂ，Ｇｃを取得し、次の式に従って、ＧＰ行列を求める（ステップ２５）。

ＧＰ＝Ｇａ・Ｇｂ・Ｇｃ

従って、本実施形態におけるＧＰ行列は、Ｇａ行列の次元合わせを行った行をそのまま行にとり、ユーザーの各メンバーを列にとってなっており、ＧＰ行列の各要素は、メンバー毎の過去の文書処理における処理重要語句の処理重要度ｆ（ｘ）に各メンバーの重要度を加味して表した数値となっている。

【００３１】ＧＰ行列が取得されると、続いてＣＰＵ１１１は、このＧＰ行列をもとにＧＰベクトル（嗜好ベクトル）を取得する（ステップ２６）。図６は、ＧＰ行列からＧＰベクトルを算出する行程を概念的に説明する説明図である。

【００３２】ＣＰＵ１１１は、まず、ＧＰ行列の各要素ｇｉｊ（ｉ＝１～メンバー数ｍ、ｊ＝１～処理重要語句の和集合の数ｋ）の各行毎の要素の平均値を算出して列ベクトル（総ＧＰベクトル）を得る（図６（１）→（２））。この総ＧＰベクトルは、各要素ｇｉが処理重要語句毎のユーザーグループ全体における過去の文書処理での出現頻度（但し各処理重要語句の予め決められた処理重要語句の重み等や、メンバーの重要度が加味されている）を反映した数値となっている。ＣＰＵ１１１は、更に、この総ＧＰベクトルの各要素ｇｉを文書の処理回数の総数で割って、１列のＧＰベクトルを得る（図６（２）→（３））。この様に、総ＧＰベクトルを文書の処理回数の総数で割るのは、行列Ｇｂに文書の処理回数が要素として含まれており、処理回数が増えるに従ってＧＰベクトルが大きくなっていくのを回避するためである。

【００３３】そして、ＣＰＵ１１１は、ＧＰベクトルの各要素とこの各要素に対応する文書ベクトルＶの要素と

を掛け合わせて、嗜好文書ベクトルV'を得る。嗜好文書ベクトルV'は、嗜好文書ベクトルデータベース１６７に格納して（ステップ２６）、嗜好文書ベクトル取得処理を終了する。

【００３４】図７は、文書ベクトルを嗜好文書ベクトルにシフトさせた状態を概念的に示す説明図である。尚、この説明図では、表示の都合上処理重要語句をX，Yの２つとして説明するが、処理重要語句の数が異なっていても、文書ベクトルを嗜好文書ベクトルにシフトさせた状態については本質的に同様である。図７に示すように、文書ベクトルＶｐ＝（０，１）、文書ベクトルＶｑ＝（２，１）、及び文書ベクトルＶｒ＝（８，１）をＧＰベクトル＝（１，２）によりシフトさせたものである。文書ベクトルＶｐ、文書ベクトルＶｑ、文書ベクトルＶｒは、ＧＰベクトルにより、それぞれ嗜好文書ベクトルＶｐ'＝（０，２）、Ｖｑ'＝（２，２）、Ｖｒ＝（８，２）にシフトされる。このように、出現頻度等によって決定される要素値により、文書ベクトルから嗜好文書ベクトルへ様々な角度でシフトされる。

【００３５】次に、本実施形態による類似度算出処理について図８を参照して説明する。図８は、類似度算出処理の動作を示すフローチャートである。

【００３６】類似度算出に際しては、ＣＰＵ１１１は、類似度を算出する基準となる文書（基準文書Ａ）についての文書嗜好ベクトルV'ａを取得する（ステップ２１）。基準文書Ａについての文書嗜好ベクトルV'ａは、上述の動作によって、または既に嗜好文書ベクトルデータベース１６７に格納されている場合にはこの文書嗜好ベクトルデータベース１６７から読み出して、取得する。尚、基準文書Ａは、文書データベース１６４に格納されている文書等から処理時に選択しても、予め決定されているものを用いてもよい。また、基準文書Ａとしては、１つの文書の他、複数の文書よりなる文書群や、文書群からクラスタリング処理により自動抽出した文書群を用いることもできる。

【００３７】また、基準文書Ａとの類似度を算出する比較文書Ｂについての文書ベクトルＶｂを取得する（ステップ２２）。

【００３８】次に、ＣＰＵ１１１は、基準文書Ａと比較文書Ｂとの類似度Ｓｂを、基準文書Ａの嗜好文書ベクトルＶａ'と比較文書Ｂの文書ベクトルＶｂ間の角度ｑに依存するコサインにより求める（ステップ２３）。すなわち、比較文書Ｂの基準文書Ａに対する類似度Ｓｂは、嗜好文書ベクトルＶａ'と文書ベクトルＶｂの間の角度をｑ、嗜好文書ベクトルＶａ'と文書ベクトルＶｂの内積をＶａ'・Ｖｂ、嗜好文書ベクトルＶａ'，文書ベクトルＶｂの大きさをそれぞれ｜Ｖａ'｜、｜Ｖｂ｜とした場合、次の数式１により求まる。

【００３９】

【数式１】類似度Ｓｂ＝ＣＯＳ（ｑ）＝（Ｖａ'・Ｖｂ）／（｜Ｖａ'｜×｜Ｖｂ｜）

【００４０】この類似度Ｓｂの値は－１≦Ｓｂ≦１までの値をとり、１に近いほど比較文書Ｂの文書ベクトルＶｂと基準文書Ａの嗜好文書ベクトルＶａとの向きが近く、文書ベクトルＶｂが嗜好文書ベクトルＶａに類似し、比較文書Ｂが、ユーザーの嗜好に近いと考えることができる。

【００４１】ＣＰＵ１１１は、求められた類似度Ｓｂを比較文書Ｂと対応させて類似度格納エリア１１３７に格納して（ステップ２４）、類似度算出処理を終了する。

【００４２】この様に、本実施形態では、ユーザーの処理文書中における処理重要語句の頻出頻度をもとに、基準文書Ａの文書ベクトルVをシフト（文書ベクトルの各要素値を変換）してユーザーの嗜好を加味した嗜好文書ベクトルV'を取得し、この嗜好文書ベクトルV'に対する他の文書の類似度が算出される。従って、この類似度は、ユーザーの興味や注目度、目的等の嗜好ユーザーの嗜好に対する指標となるので、この類似度に基づいて文書の分類や検索を行うことにより、ユーザーの嗜好を反映した分類や検索が可能となる。また、ユーザーの興味や注目度等に合う文書を選択して配信することが可能となる。本実施形態によると、ＧＰ行列を用いた変換によって文書ベクトルVを嗜好文書ベクトルV'にシフトさせているので、計算処理が簡単であり、ベクトル空間法を採用したコア・エンジンを備えた一般の文書処理装置に容易に適用することが可能である。

【００４３】本実施形態よると、文書ベクトルを嗜好文書ベクトルにシフトさせるＧＰベクトル（嗜好ベクトル）のもととなるＧＰ行列を、表現すべき特徴毎の単純な観点で構成した行列Ｇａ，Ｇｂ，Ｇｃの掛け合わせて求めているので、様々な特徴を考慮に入れたＧＰ行列を容易に構成して文書ベクトルVをシフトさせることが可能である。本実施形態よると、文書ベクトルを嗜好文書ベクトルにシフトさせるＧＰベクトルのもととなるＧＰ行列は、各列がユーザーのメンバーの興味を反映しているので、ユーザーを数グループに分割した該グループのＧＰ行列や個々のメンバーのＧＰ行列（ベクトル）を容易に得ることができる。本実施形態よると、ＧＰ行列がユーザーの過去に処理した文書をもとに適宜書き換えられている行列Ｇａ，Ｇｂ，Ｇｃをもとに取得されているので、文書ベクトルVがユーザーの嗜好の経時的掲示変化に対応した嗜好文書ベクトルV'にシフトされ、ユーザーの嗜好の変遷に追随した類似度の算出および検索・分類等の処理が可能となる。

【００４４】尚、本発明は、上述の実施形態に限定されるものではなく、本発明の趣旨を逸脱しない限りにおいて適宜変更が可能である。上述の実施形態においては文書処理装置としてコンピュータを用いているが、コンピュータに限定されるものではなく、ワードプロセッサ等であってもよい。上述の実施形態においてはＧＰ行列

は、メンバー毎の過去の文書処理回数（行列Ｇａ）と各文書における処理重要語句の出現頻度（行列Ｇｂ）、および各メンバーの重要度（行列Ｇｃ）とから取得されているが、メンバー毎の過去の文書処理回数（行列Ｇａ）と各文書における処理重要語句の出現頻度（行列Ｇｂ）のみにより取得されてもよい。また、例えば、各文書の処理時間や、他の文書作成に引用された件数、リンク付けされている数等も加味して取得されてもよい。更に、ＧＰ行列を上述の実施形態と同様に行列Ｇａ～行列Ｇｃ等の行列の掛け合わせから取得する場合において、行列Ｇａ～行列Ｇｃ等の各行列の要素はそれぞれ処理重要語句の文書中の出現頻度や、メンバーが各文書を処理した回数を反映した数値となっていればよく、直接出現頻度や処理回数そのものを表していなくてもよい。

【００４５】上述の実施形態においては行列Ｇａ～Ｇｃは過去の文書処理内容から取得されているが、ユーザーが取得して行列データベース１６８に直接入力してもよい。上述の実施形態においては行列Ｇａ～Ｇｃは所定期間毎に書き換えられているが、文書処理を行う毎に、または所定回数の文書処理を行う毎等に書き換えてもよい。ＧＰベクトルを表示装置に表示するＧＰベクトル表示手段を備え、ユーザーやユーザーメンバーの嗜好を視覚的に把握できるようにしてもよい。この場合、ＧＰベクトルを行列データベースまたは専用のＧＰベクトルデータベースに経時順に格納しておき、経時変化も把握できるようにしてもよい。

【００４６】説明した実施形態では処理重要語句や処理重要度を取得する手法として図４のフローチャートに従った方法を１例にして説明したが、本発明でこの方法に限られるものではなく、文書中から処理重要語句を抽出する方法や、処理重要度の決定方法等については、公知の各種方法により置き換えることが可能である。更に、２つの文書嗜好ベクトルの類似度の算出方法については、数式１により類似度を算出することとしたが、この数式に限定されるものではなく、文書嗜好ベクトル相互間の類似関係を表すことが可能であれば他の数式により類似度を算出することも可能である。算出した類似度の表示は、類似度の操作者からの入力により類似閾値を取得し、当該類似閾値よりも高い類似度を備えた対象文書のみを表示させたり、類似度の高いうちから１０個の文書のみを表示させたりすることもできる。また、類似度の高い順ではなく、あいうえお順等に表示された対象文書名とともに表示してもよい。更に、類似度表示は、操作者からの命令のあったときのみに表示させるようにしたり、表示装置には表示させずに印刷させることとしてもよい。

【００４７】説明した実施形態は日本語で作成された文書に限られるものではなく、あらゆる言語で作成された文書を対象とすることが可能である。その場合、対象となる文書が作成された言語用の形態素解析アルゴリズム等を使用するといった、本発明の構成には影響のない部分を変更するだけでよい。

【００４８】なお、以上の実施形態において説明した、各装置、各部、各動作、各処理等に対しては、それらを含む上位概念としての各手段（～手段）により、実施形態を構成することが可能である。例えば、「文書データベース１６４から、各処理文書の処理回数を取得し（ステップ１４）」との記載に対して文書の処理回数を記憶する処理回数データベースを文書データベース１６４とは別途に構成したり、「処理回数取得手段」を構成したり、「抽出した候補語（句）の処理文書での出現頻度、評価関数から、各候補語（句）の処理重要度ｆ（ｘ）を取得する（ステップ１２３）との記載に対して、「処理重要語句取得手段」を構成するようにしてもよい。同様に、その他各種動作に対して「～（動作）手段」等の上位概念で実施形態を構成するようにしてもよい。

【００４９】

【発明の効果】以上説明したように、本発明によれば、嗜好文書ベクトル取得手段により文書ベクトルをユーザーの嗜好を加味した嗜好文書ベクトルにシフトさせ、この嗜好文書ベクトルに対する類似度を取得することにより、ユーザーの興味や注目度、目的等の嗜好に対する文書の類似度が取得でき、この類似度に基づいて分類や検索を行うことにより、ユーザーの興味や注目度、目的等の嗜好を反映した分類や検索、配信等の文書処理が可能となる。

【図面の簡単な説明】

【図１】本発明の文書処理装置の一実施形態であり、本発明の文書処理プログラムが記憶された記憶媒体の一実施形態の該プログラムが読み取られたコンピュータの構成を表したブロック図である。

【図２】図１の実施形態における行列Ｇａ，Ｇｂ，Ｇｃを示す説明図である。

【図３】図１の実施形態による行列Ｇａ，Ｇｂ書き換え処理の動作を表したフローチャートである。

【図４】図１の実施形態による処理重要語句・処理重要度取得処理の動作を表したフローチャートである。

【図５】図１の実施形態による嗜好文書ベクトル取得処理の動作を示すフローチャートである。

【図６】図１の実施形態におけるＧＰ行列からＧＰベクトルを算出する行程を概念的に説明する説明図である。

【図７】図１の実施形態における文書ベクトルを嗜好文書ベクトルにシフトさせた状態を概念的に説明する説明図である。

【図８】図１の実施形態による類似度算出処理の動作を示すフローチャートである。

【図９】請求項１に記載した発明のクレーム対応図である。

【図１０】請求項３に記載した発明のクレーム対応図である。

15　　16

【図１１】請求項５に記載した発明のクレーム対応図である。
【図１２】請求項６に記載した発明のクレーム対応図である。
【図１３】請求項８に記載した発明のクレーム対応図である。
【図１４】請求項１０に記載した発明のクレーム対応図である。
【図１５】請求項１１に記載した発明のクレーム対応図である。
【符号の説明】

１１	制御部
１１２	ＲＯＭ
１１３	ＲＡＭ
１１３１	対象文書格納エリア
１１３２	キーワード格納エリア
１１３４	文書ベクトル格納エリア
１１３５	行列格納エリア
１１３６	嗜好文書ベクトル格納エリア
１１３７	類似度格納エリア
１２	キーボード
１３	マウス
１４	表示装置
１５	印刷装置
＊１６	記憶装置
１６１	仮名漢字変換辞書
１６２	プログラム格納部
１６４	文書データベース
１６５	重要語データベース
１６６	文書ベクトルデータベース
１６７	嗜好文書ベクトルデータベース
１６８	行列データベース
１７	記憶媒体駆動装置
１８	通信制御装置
１９	入出力Ｉ／Ｆ
１０１	嗜好ベクトル取得手段
１０２	文書取得手段
１０３	文書ベクトル取得手段
１０４	シフト手段
１０５	ＧＰ行列取得手段
１０６	類似度算出手段
２０１	嗜好ベクトル取得機能
２０２	文書取得機能
２０３	文書ベクトル取得機能
２０４	シフト機能
２０５	ＧＰ行列取得機能
２０６	類似度算出機能

＊

【図２】

(a)　行列　Ga（キーワード，文書）

	文書A	文書B	文書C…
重要	2	1	18 …
重要語	5	18	0 …
重要度	1	19	0 …
政治	2	0	21 …
⋮	⋮	⋮	⋮

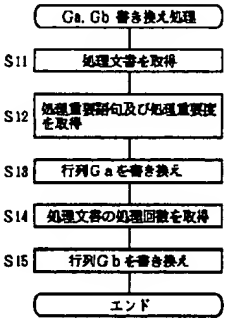
(b)　行列　Gb（文書，処理者）

	皇太郎	花園美子	黒帯三四郎	意見五郎
文書A	1	0	1	0
文書B	1	1	2	0
文書C	1	1	1	1
⋮	⋮	⋮	⋮	⋮

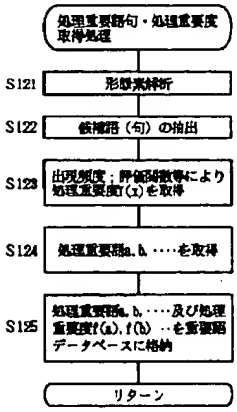
(c)　行列　Gc（処理者の重要度）

	皇太郎	花園美子	黒帯三四郎	意見五郎
皇太郎	1.5	0	0	0
花園美子	0	0.8	0	0
黒帯三四郎	0	0	1.3	0
意見五郎	0	0	0	1.1

【図３】

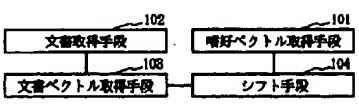


【図４】

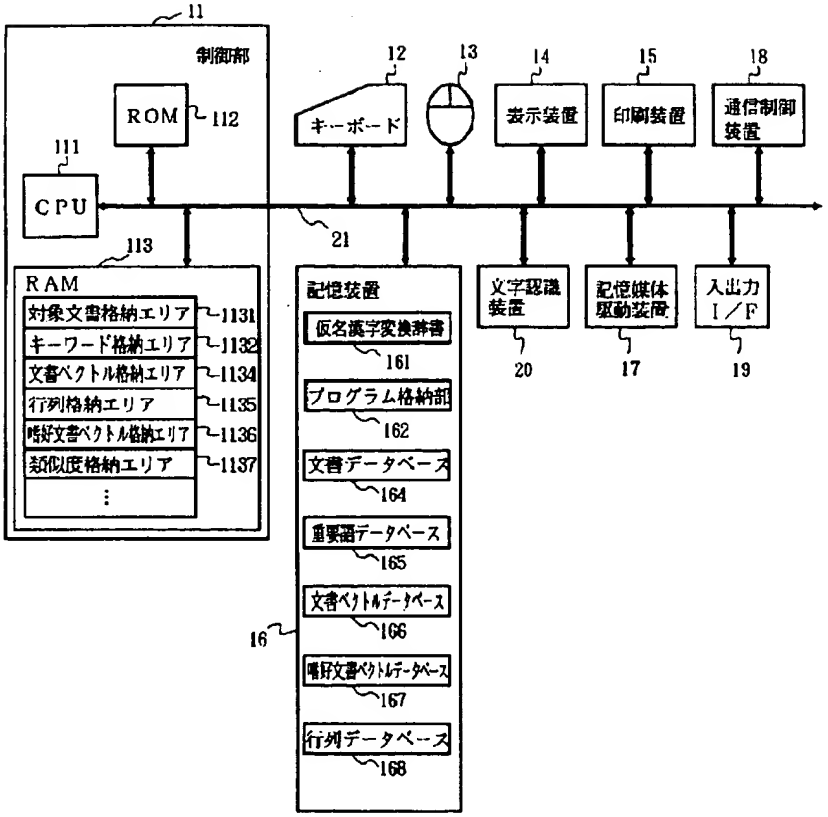


【図５】

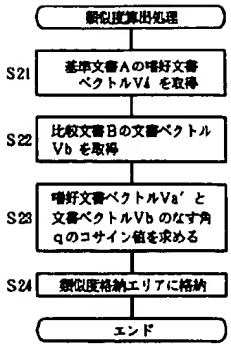
【図９】



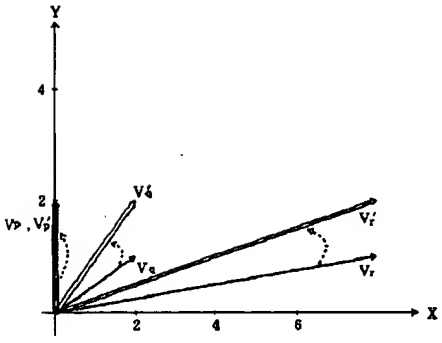
【図１】



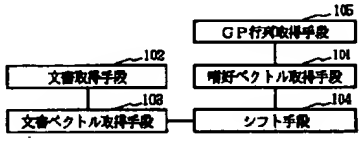
【図８】



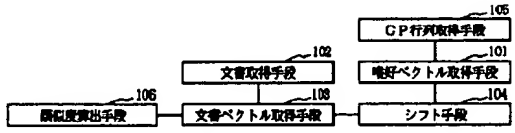
【図７】



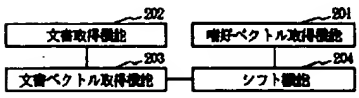
【図１０】



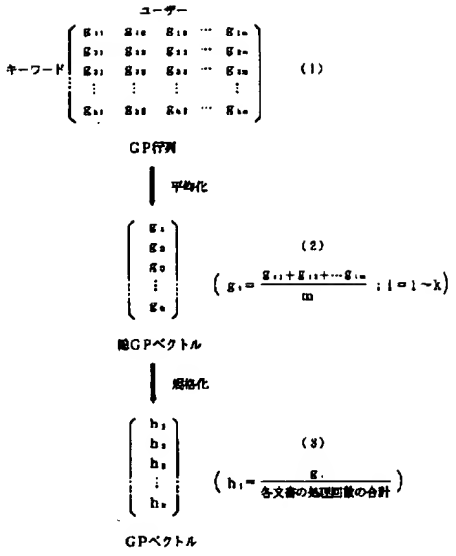
【図１１】



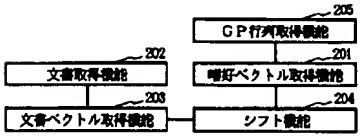
【図１２】



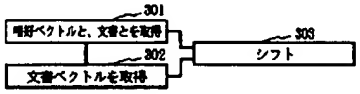
【図６】



【図１３】



【図１５】



【図１４】